I/O
January-March, 2014
A quarterly newsletter for LANL's HPC user community

In this issue

LA-UR-14-21461

Things you should know.....

Standard service features

The Integrated Computing Network (ICN) Consulting Team provides user support on a wide variety of HPC topics:

- Programming, languages, debugging
- Parallel computing
- HPC and Unix operating systems, utilities, libraries
- Unix / Linux scripting
- Archival storage: High Performance Storage System (HPSS), General Parallel File System (GPFS)
- Desktop backup (TSM), storage, file transfer, network
- Mercury: Cross-network file transfer service
- HPC infrastructure in both the Open and Secure

Service hours

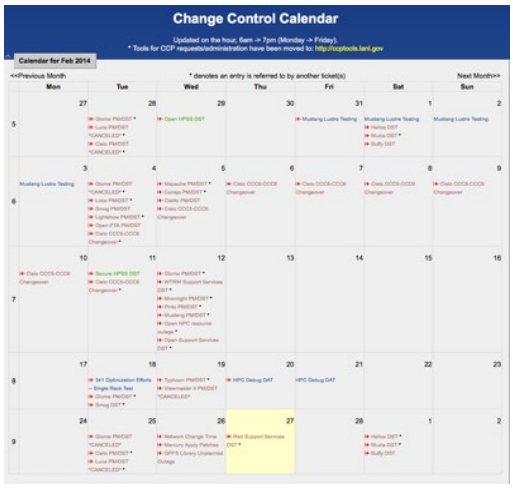
Support available Monday through Friday, 8 a.m. - 12 p.m. and 1 p.m. - 5 p.m. After-hours callers have the option to transfer to our 7x24 computing facility operations desk. Please use this option to report any situations requiring urgent after hours support. Email or voicemail messages may be left 24 hours/day. We will respond as quickly as possible during normal hours.

Phone #
505-665-4444 opt 3

Email
consult@lanl.gov

Documentation
http://hpc.lanl.gov

HPC Change Control Calendar
http://ccp.lanl.gov



Consultants Corner

Moab Job Priority

One of the most frequent questions we receive from customers is, "why is my job not running?" Those who take the time to ask usually want more information than, "the cluster is busy and your job has to wait."

Usually the answer is related to the Moab scheduler trying to accommodate all of its submitted jobs. Frequently it is a matter of just waiting for your job's turn to run, but we can provide much more information about job scheduling for users who wish to know more and want to understand the job priority calculation.

On the LANL HPC clusters, we use Priority Group scheduling on our ASC platforms and Moab Fairshare scheduling on our IC platforms. You can look up the policy for any of our clusters here: http://hpc.lanl.gov/scheduling_policies.

We explain LANL HPC job scheduling parameters for all of our platforms here and are happy to schedule a presentation for your Group or Team: http://hpc.lanl.gov/files/Moab_Priority.pdf

We provide more detail on how Fairshare calculates your job priority, and also share a set of commands that we use to understand priorities of specific jobs here: http://hpc.lanl.gov/moab_fairshare

Of course, there may be other reasons why your job is not running; it might be stuck in the queue waiting on resources that will never be available (such as an unattainable PPN value), it might depend on a previous job or it might be blocked. When investigating this issue, our first question is answered by the Moab command: "checkjob -v <JobID>". The last couple of lines from its report will offer clues. You can find interpretations of results from this command here: http://hpc.lanl.gov/moab_idiosyncrasies

And, as always, if you are not able to answer your question to your satisfaction, please contact ICN Consulting: consult@lanl.gov, or 505-665-4444 option 3.

Consultants
left to right, back to front
Ben Santos, Hal Marshall, Riley Arnaudville,
Rob Derrick
Giovanni Cone, Rob Cunningham
David Kratzer

Software and Tool News

The Programming Environment and Runtime Team, formerly PTools, of LANL's High Performance Computing Division, strive to provide useful tools and knowledgeable staff to assist users in debugging and optimizing applications on the production clusters. Some of the services they provide include one-on-one developer support, hosting workshops with third-party software vendors, user educational outreach, and software documentation hosted on hpc.lanl.gov

February TotalView Workshop:

In February, the team hosted a Rogue Wave Software workshop to demonstrate and provide hands-on assistance using TotalView, a defect and memory leak analysis tool provided on our production clusters.

The representatives updated participants on forward-looking projects, such as debugging at large scale (~800K cores test) and heterogeneous architectures. TotalView is undergoing changes to support parallel application debugging in these kinds of environments. In particular, they are looking at how to best support GPGPUs, and Intel Xeon Phis, among other technologies. They also showed how to do memory debugging and the ability to use TV's Deterministic Replay Engine to debug backwards, from a crash point to the source-code location that may have caused the error.

Introductory videos can be viewed at:
http://www.roguewave.com/products/totalview/resources/videos.aspx

LANL, in partnership with Rogue Wave, Livermore National Lab, and Sandia National Lab, continue to improve upon TotalView's scalability and usability through a tri-lab contract. This work is ongoing and we are always looking for user feedback and improvement suggestions.

Darshan: A Scalable HPC I/O Characterization Tool
Darshan is an I/O performance analysis tool available to users to collect and summarize valuable parallel I/O performance statistics during application runtime.

To capture I/O behavior during an application's run, load an MPI library (OpenMPI 1.6.3+, or any Mvapich2 production builds), then load the Darshan runtime modulefile. Finally, set the DARSHAN_LOG_PATH environment variable to a directory that the Darshan log file can be written to:

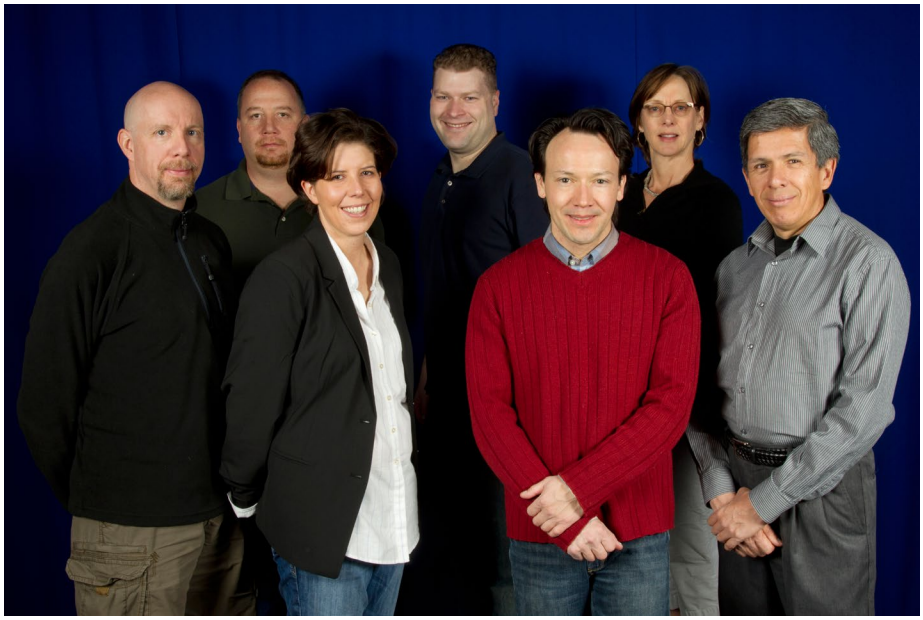
```
$> module load friendly-testing
$> module load <openmpi|mvapich2>
$> module load darshan-runtime
$> export DARSHAN_LOG_PATH=</path/to/dir> # for bash OR
$> setenv DARSHAN_LOG_PATH </path/to/dir> # for tcsh
$> ## Run application as you normally do once environment is set up for Darshan
$> module load darshan-util
$> darshan-job-summary.pl $DARSHAN_LOG_PATH/log.file
$> ## This script will generate a report of i/o usage of your app at runtime, in
$> ## .pdf; alt output options are possible http://hpc.lanl.gov/content/darshan
```

If you suspect I/O is a bottleneck in the runtime of your application, or you're curious about the performance differences in various runtime options for your application, give Darshan a try. Please report any feedback or questions about Darshan to: ptools@lanl.gov and check out the LANL documentation pages regarding Darshan: http://hpc.lanl.gov/content/darshan

HPC Technical Sessions

The second week of February 2014, the HPC Division launched a monthly initiative to support HPC users in an open forum. This is an excellent venue to bring runtime problems, ask questions, and get one-on-one consultation with HPC staff. The nature of these sessions is to extend the services already provided by the Consultants and the PE&R team. The intent is to make available our resources in an unstructured format in order to address any barriers to success that users may encounter while running on HPC production clusters at LANL. We plan to tackle big-picture development issues with porting code to specialty systems or provide assistance in running tools provided on our clusters. The Tech Sessions' success will be highly dependent on user feedback and participation. Stay tuned for the next workshop announcement via ICN Consulting. We look forward to rolling up our sleeves to help you so that you may achieve quicker success in your scientific endeavors on LANL HPC resources!

As always, the Programming Environment and Runtime Team of HPC-3 is dedicated to providing excellent customer support and a quality Programming Environment to the users of our High Performance Computers. If you have any concerns or feedback to provide, please feel free to contact us at ptools@lanl.gov.



Programming Environments and Runtime Team
left to right
David Gunter, Riley Arnaudville, Jennifer Green, David Shrader, Giovanni Cone, Marti Hill, and Jorge Roman

Machines (coming/going/planned/technologies)

LANL's Newest HPC System (Wolf)
LANL's Institutional Computing (IC) is in the process of purchasing a new machine for its IC users. Wolf will consist of 616 compute nodes with 9856 Intel Sandy Bridge processors and an Intel/Qlogic interconnect. This new machine will reside in the Turquoise partition.

Wolf is a new general purpose computing cluster with the same general architecture as Pinto (in the open) and Luna (in the secure). It will be 4 times the size of Pinto or 4/10 the size of Luna. The system is manufactured by Cray (formerly Appro) and consists of 616 compute nodes each with 2 Intel Xeon E5-2670 (Sandy Bridge EP) 8-core CPUs running at 2.6 GHz and 64 GB of memory (4 GB per core). In aggregate, the system will have 9,856 compute processor cores and provide 197 TF per sec of computational capability. The machine interconnect is a traditional Infiniband QDR fat tree. Wolf will be able to access both our existing Panasas parallel file system as well as the new Lustre parallel file system.

Wolf

Wolf's home will be in LANL's Turquoise open collaborative computing environment. As with all machines in the Turquoise, the system is shared between programs contributing to the network (Institutional Computing, Climate, and ASC). Each program manages access separately; LANL's Instututional Computing Program holds an annual Call For Proposals for access. The current call was announced in late December, with proposals due February 12.

With the arrival of Wolf we will retire the aging Lobo system. There will be also be some movement of workloads to help balance the computing environment: Lobo's "small job" workload will be moved to Pinto and Pinto's workload will move to Wolf.

For more information on Wolf, please see http://hpc.lanl.gov or contact HPC consultants at consult@lanl.gov. Wolf is expected to be operational before May 2014.

Technology Frontiers

HPC Division and Institutional Computing (IC) are fielding a new file system for the Turquoise network. The file system is based on Lustre and was provided by Data Direct Network (DDN).

The file system will provide approximately 2.6PB of new useable space and be capable of >20GBs network performance depending upon which cluster resource you are using. The file system should be open to a few friendly users in February 2014 and then generally available to all Turquoise clusters in early March 2014.

IC is also in the very early stages of deploying some long-term storage for the Turquoise network. Look for future updates about this new resource.

Advanced Technology Futures

HPC division in cooperation with "Alliance for computing at Extreme Scale" (ACES LANL/SNL partnership) is in the early procurement stages to acquire the first NNSA/ASC Advanced Technology System (ATS-1). This machine, named Trinity, will embody new architectural concepts to enable much faster check-point and recovery mechanisms. Look for future announcements about this exciting new resource.

Big Data at LANL: Hadoop on Glome and Kugel

Over the last several months, Reese Baird and Tim Randles (HPC-3) have been building two big data clusters out of old Redtail hardware. These systems are named Glome and Kugel, and will initially run the Hadoop map/reduce implementation, to give laboratory scientists a user testbed to work with these technologies that are commonly used in the world of big web companies.

Glome currently consists of 96 nodes, each with 8 CPUs, 32GB of RAM and 8TB of disk. Apache Hadoop 2 is installed across the whole system and is being managed with the Yarn resource manager; the system has a 687TB HDFS filesystem spread across it. The system is available for early users who can tolerate changes in the software stack and downtimes for experimentation or reconfiguration as needed. The HPC consultants are interested in early users who want to offer constructive feedback on the current hardware and software stack, and how the stack should change or stay the same on future big data systems. Accounts can be requested at http://hpcaccounts.lanl.gov/.

Here is a presentation on Hadoop and Glome given by Tim Randles.

Tim Randles with Glome

HPC- Behind the Scenes

The Future LANL HPC Environment, a Story of Change

In the seven decades of computing at LANL, there have been many generations of computing capabilities. It wasn't until the 1960's that an integrated computing environment, with storage, input, output, networking, and computing emerged. Eventually this environment became known as the Integrated Computing Network. As the HPC environment at LANL moves toward the future, it is important to have a perspective of the past.

The future of high performance computing seems to be dominated by everything but computing. Power, cooling, rack weight, data management, and other infrastructure related items have grown, as a total cost of computing, from far less than 10% a decade ago, to well over 25% now, and that number is rapidly growing.

The entire story of the LANL HPC environment is a fascinating exploration into how HPC has grown over the years.

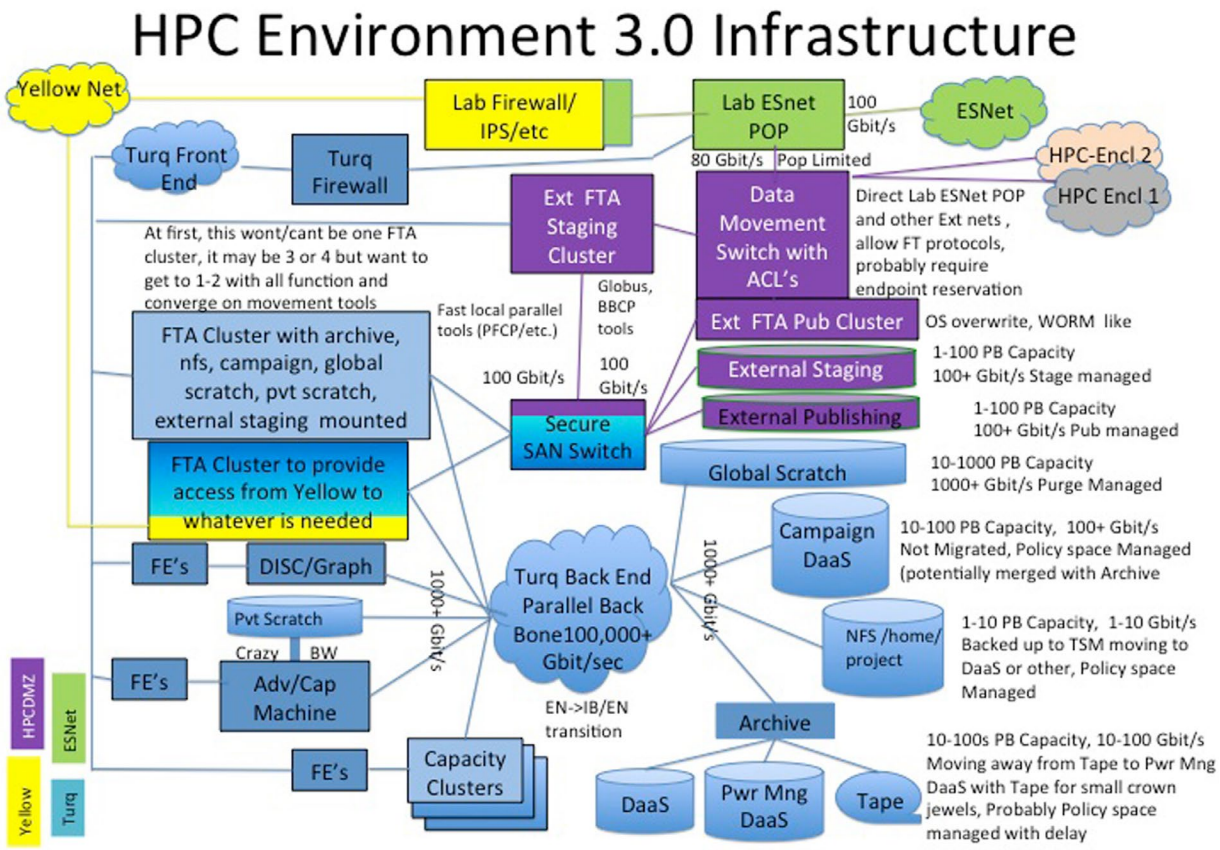
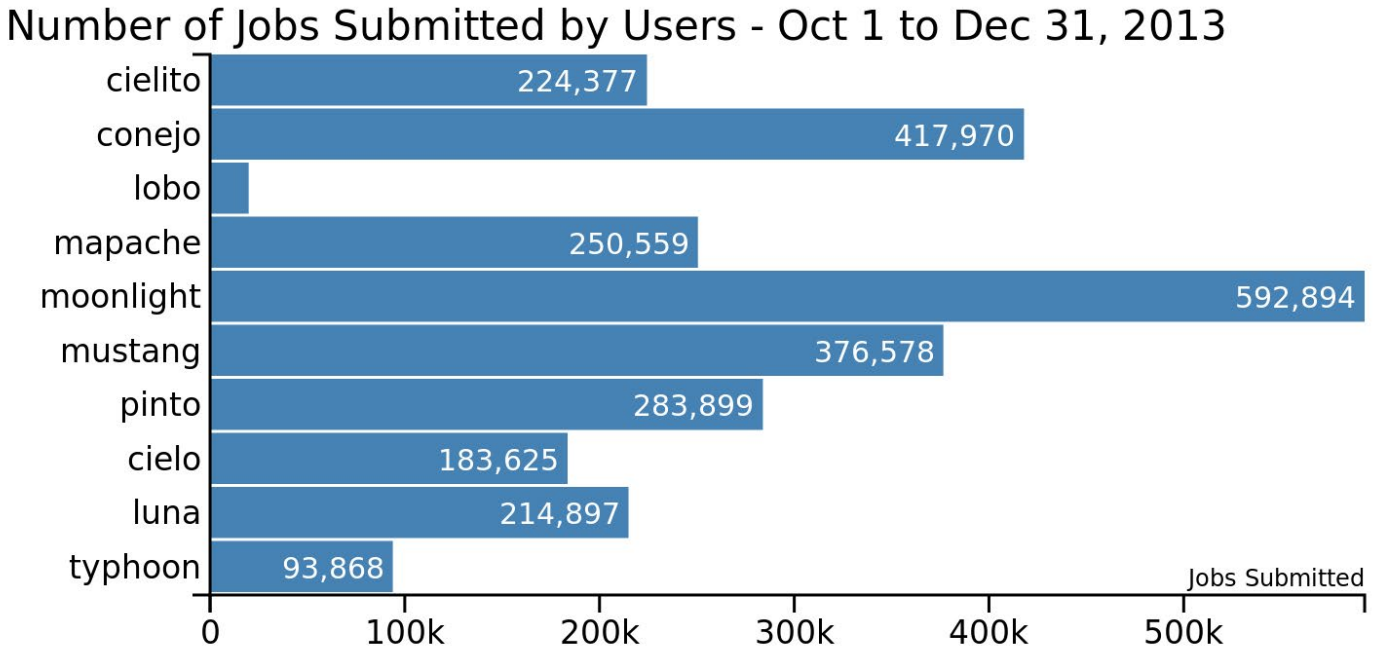
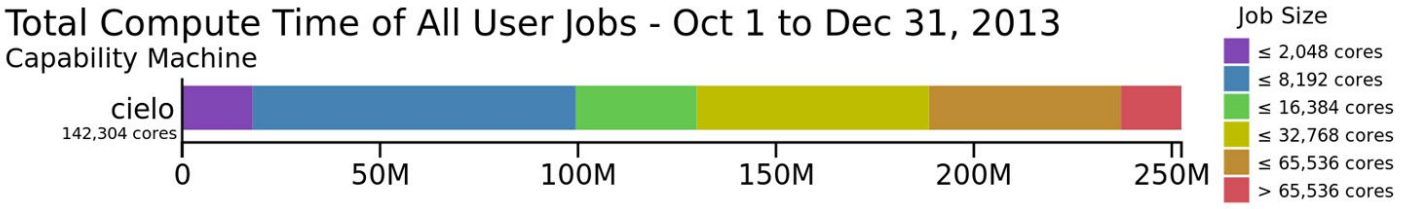
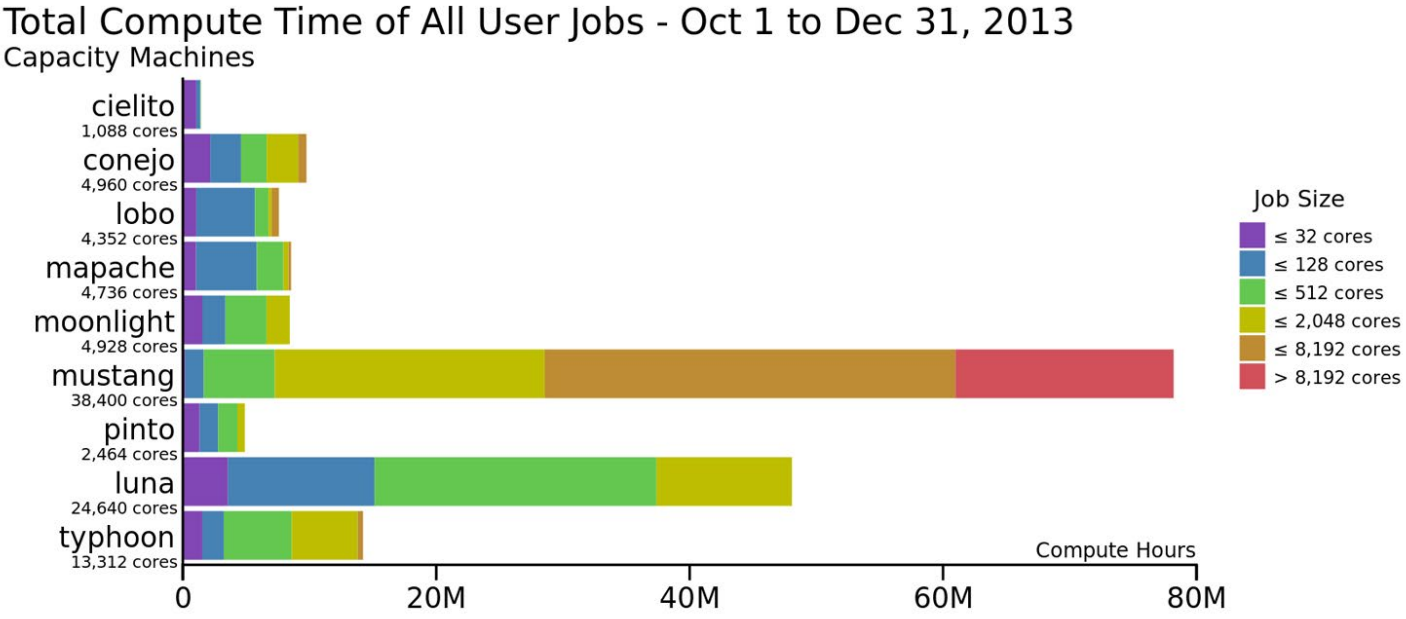


Figure 1: The diagram is a notional diagram of the future open computing environment in the 2014-2017 timeframe.

Quarterly Statistics



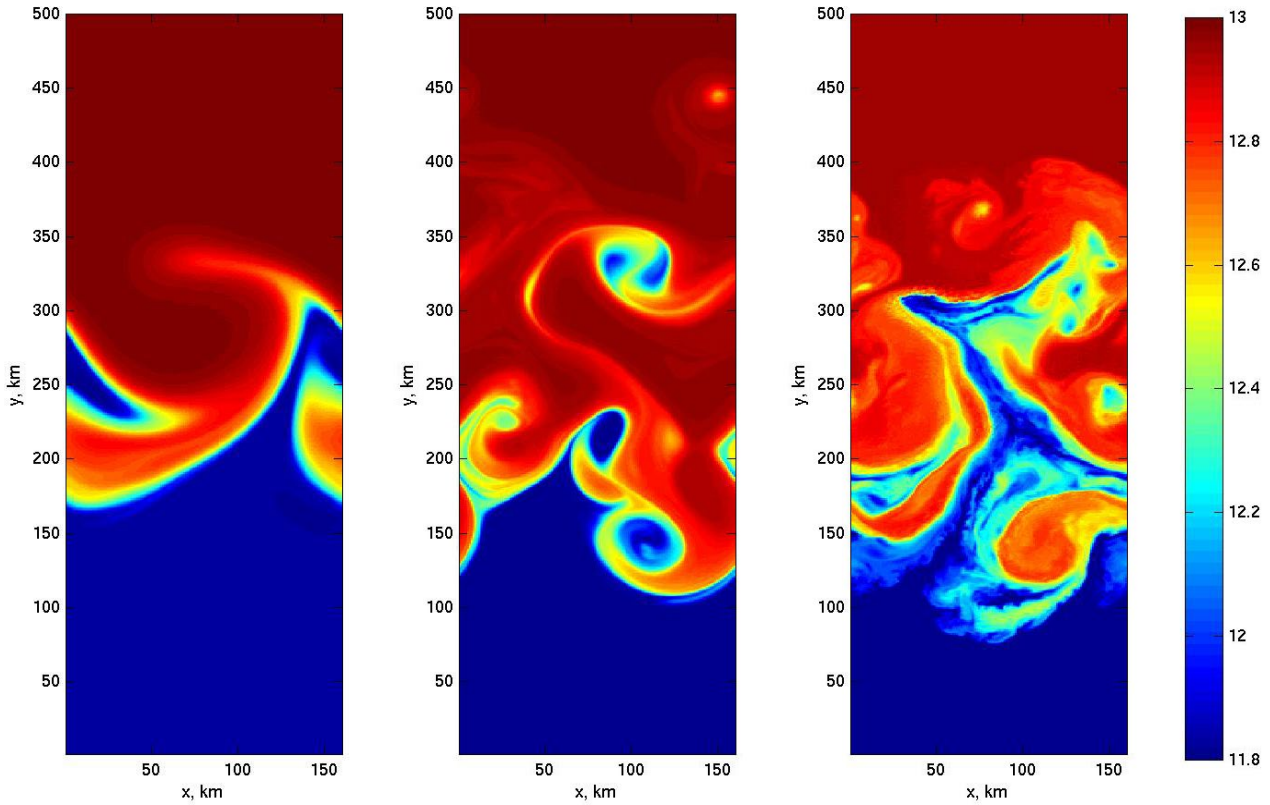




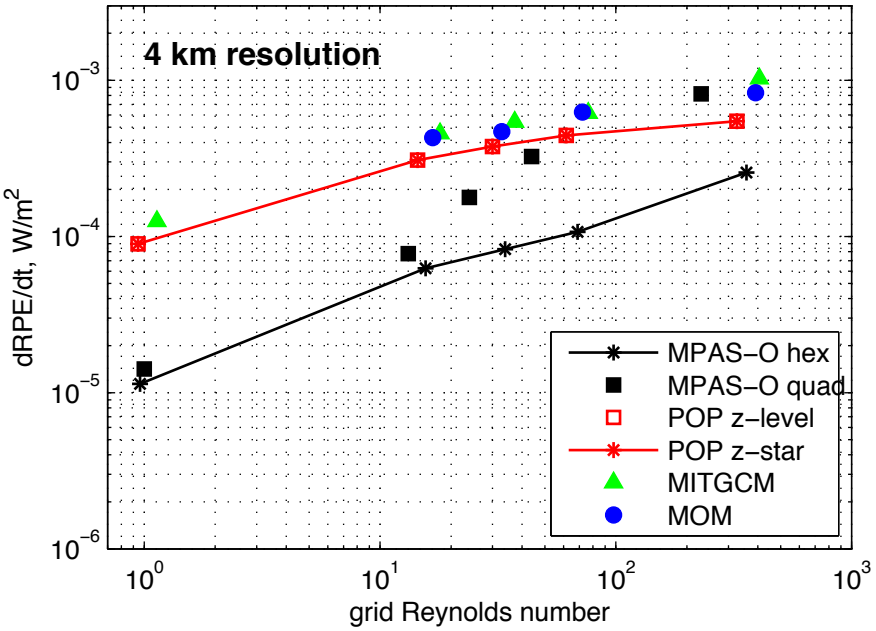
HPC user Highlight

MPAS-Ocean (Model for Prediction Across Scales), is a new open source ocean-climate model developed by the COSIM (Climate, Ocean, and Sea Ice Team) at LANL (CCS-2 and T-3). Developed by Mark Peterson, Todd Ringler, and Douglas Jacobsen (T-3), it was released in June 2013 and is available at http://mpas-dev.github.io. MPAS-Ocean uses a horizontally unstructured mesh based on Voronoi tessellations that enable regional high-resolution grids within a global grid, with smooth transitions in between.

In the development of a new ocean-climate model, verification of model behavior against observations and standard test cases is a critical part of the development process. One of six test cases is the baroclinic channel, which is an idealization of the Southern Ocean, and is used to measure fluid mixing in the presence of eddies (Figure 1).



A parameter study is run for each test case, varying viscosity and horizontal resolution. Statistics are compared directly to long-standing ocean models (Figure 2).



The MPAS-Ocean model was verified with hundreds of simulations on Mustang, Lobo, and Conejo using Institutional Computing resources. These ranged from single-node jobs at low resolution to 256 node (6144 core) jobs on Mustang for global high-resolution simulations. My advice to HPC users is to plan ahead for large parameter studies. I give a numbered ID to each simulation, track them in the rows of a spreadsheet with parameter values in the columns, and use corresponding names for run directories on the scratch drives. Text tools like sed and awk within shell scripts, can automate the creation of input decks for large parameter studies. For large jobs, I/O is often the bottleneck. MPAS uses PIO and pNetcdf libraries to write large files in parallel, and reduce the number of cores involved in I/O to one per node.

Current Machines- A snapshot in time

Name (Program[1])	Processor	OS	Total Compute Nodes	CPU cores per Node/ Total CPUs	Memory per compute Node/Total Memory	Interconnect	Peak (TFlop/s)	Storage
Secure Restricted Network (Red)								
Cielo (ASC)	AMD Magny-Cours	SLES-based CLE and CNL	8,894 nodes	16/142,304	32 GB/297 TB[5]	3D Torus	1,370	10 PB Lustre
Luna TLCC2 (ASC)	Intel Xeon Sandybridge	Linux (Chaos)	1540 nodes	16/24,640	32 GB/49 TB	Qlogic InfiniBand Fat-Tree	513	3.7 PB Panasas
Typhoon (ASC)	AMD Magny-Cours	Linux (Chaos)	416 nodes	32/13,312	64 GB/26.6 TB	Voltaire InfiniBand Fat-Tree	106	3.7 PB Panasas
Open Collaborative Network (Turquoise)								
Cielito (ASC)	AMD Magny-Cours	SLES-based CLE and CNL	68 nodes	16/1088	32 GB/2.3 TB[5]	3D Torus	10.4	344 TB Lustre
Conejo (LC)	Intel Xeon x5550	Linux (Chaos)	620 nodes	8/4960	24 GB/14.9 TB	Mellanox Infiniband Fat-Tree	52.8	1.8 PB Panasas
Lightshow[3] (ASC)	Intel Xeon	Linux (Chaos)	16 nodes	12/192	96 GB/1.5 TB	Mellanox Infiniband Fat-Tree	4.0	1.8 PB Panasas
Lobo TLCC[3] (IC)	AMD opteron	Linux (Chaos)	272 nodes	16/4,352	32 GB/8.7 TB	Voltaire Infiniband Fat-Tree	38.3	1.8 PB Panasas
Mapache (ASC)	Intel Xeon x5550	Linux (Chaos)	592 nodes	8/4736	24 GB/14.2 TB	Mellanox Infiniband Fat-Tree	50.4	1.8 PB Panasas
Moonlight TLCC2[3] (ASC)	Intel Xeon E5-2670 + NVida Tesla M2090	Linux (Chaos)	308 nodes	16/4,928 + GPUs	32 GB/9.86 TB	Qlogic Infiniband Fat-Tree	488	1.8 PB Panasas
Mustang (IC)	AMD Opteron 6176	Linux (Chaos)	1,600 nodes	24/38,400	64 GB/102 TB	Mellanox Infiniband at-Tree	353	1.8 PB Panasas
Pinto TLCC2[3] (IC)	Intel Xeon E5-2670	Linux (Chaos)	154 nodes	16/2464	32 GB/14.9 TB	Qlogic Infiniband Fat-Tree	51.3	1.8 PB Panasas

[1] Programs: IC=Institutional Computing, ASC=Advanced Simulation and Computing, R=Recharge

[3] TLCC = TriLab Linux Capacity Cluster; 2 = 2nd Generation

[5] Cielo has 372 viz nodes with 64GB memory each

[6] Cielito has 4 viz nodes with 64GB memory each